Unsupervised Machine Translation in Real-World Scenarios

**AUTHORS**
Osca de Gibert, Jakes Gorring, Urbi Aramendi-Etxeped, Olaf Perez de Villanueva, Carlos Pintor, Martina Kirchhoff-Tomé, Muxia Pérez, Gorik Labaka, Marie Mekler

**AFFILIATIONS**
Unbabel, Lisbon, Portugal
RWS Language Weaver, Dublin, Ireland
HiTZ zentroa - Ixa, UPV/EHU, Donostia, Spain
Barcelona Supercomputing Center, Barcelona, Spain

**ABSTRACT**
In this work, we present the work that has been carried out in the MT4ALL CEC project and the resources that it has generated by leveraging recent research carried out in the field of unsupervised learning. In the course of the project, 18 monolingual corpora for specific domains and languages have been collected, and 11 bilingual dictionaries and translation models have been generated. As part of the research, the unsupervised MT methodology based only on monolingual corpora ([3]) has been tested on a variety of languages and domains. Results show that in specialized domains, when there is enough monolingual in-domain data, unsupervised results are comparable to those of general domain supervised translation, and that, at any rate, unsupervised techniques can be used to boost results whenever very little data is available.

**METHOD AND RESOURCES**

**TRANSLATION MODELS**
We use Memnouns [2]

**WORD EMBEDDINGS**
We use Vecmap [3]

**BILINGUAL DICTIONARIES**
We use TermSuite

**RESULTS**

**HUMAN EVALUATION**
Three domains:
- Customer Support (ALL)
- Named Entities
- Localization: FFY-SM-DO
- Formal/Informal Registers
- Duplications: 74%:14%
- Parsing of URLs: www.comsoc.org/blog

- Financial (EN-LT)
- 100 sentences
- Mistranslations
- Named Entities
- Ottimizzazione
- Typography

- Biomedical (EN-ES)
- Full test set
- Mistranslations
- Repetitions
- Named Entities
- No-Runway errors
- MQM Score*: 150

*Reference MQM score for humans: 50

**CONCLUSIONS**
In low-resource scenarios, completely unsupervised systems tend to yield poor results, except when the amount of in-domain monolingual data is big enough to compensate.

**REFERENCES**